Testing Whether Data Follows a Power Law Distribution

24 January 2006

Eric Rasmusen

THESE ARE NOTES FOR A FUTURE PAPER, NOT A REAL PAPER YET

Abstract

A likelihood ratio test comparing spline to single parameter is a good way to test whether a distribution fits a power law.

Rasmusen: Indiana University, Kelley School of Business, BU 456, 1309 E. 10th Street, Bloomington, Indiana, 47405-1701. Office: (812) 855-9219. Fax: 812-855-3354. Erasmuse@indiana.edu. http://www.rasmusen.org.

For the most recent draft, see:
http://www.rasmusen.org/papers/netmetrics-rasmusen.pdf.

I thank Manu Raghav for helpful comments and Barick Chung for research assistance.

1. Introduction

What I will do in this note is to show how to test for a power law.

A naked-eye plot can do that.

Lehmann, B. Lautrup, and A. D. Jackson (2003) use a spline in their well-known paper on citation networks in physics. They compare with a negative exponential, I think, and cannot decide between the two.

You could use the Newman or Goldstein ML estimator in doing this.

Then find the likelihood of each of the two models and see if the log ratio flunks a chi-squared test or not.

This will show whether you have concavity or convexity instead of a power law.

Also, a one-sided test is easy.

Also, most important, you can decide whether it is close to being a power law. You have to pick your own criterion for this. I can come up with one, though. Something based on the two models estiamted, and how much they differ– the area between them in the graph. A sort of R^2, comparing within variance to between variance.

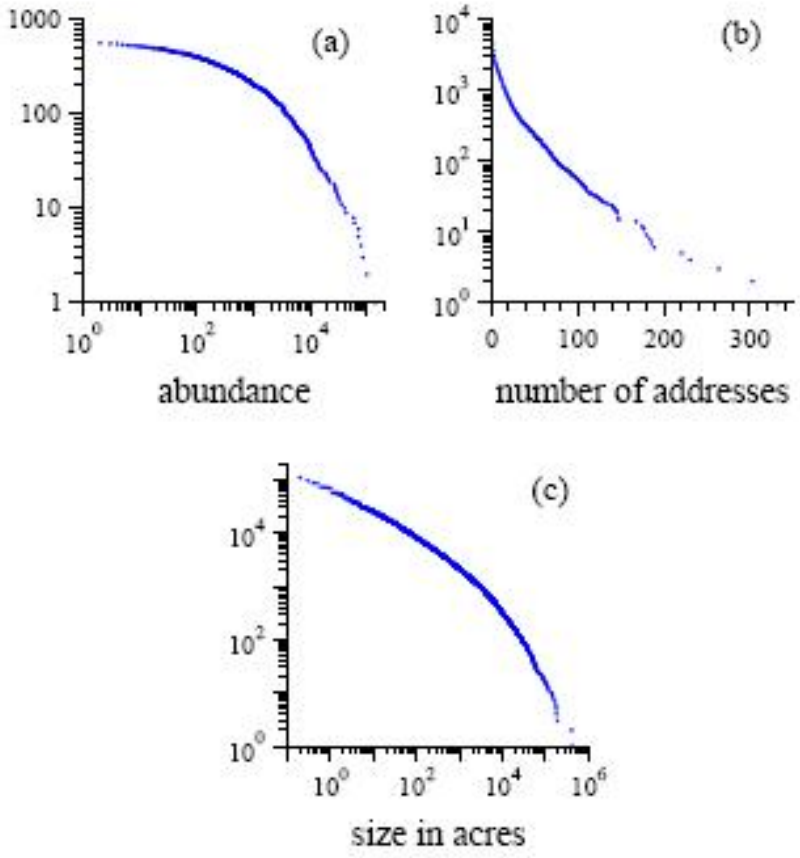

FIG. 5 Cumulative distributions of some quantities whose distributions span several orders of magnitude but that nonetheless do not follow power laws. (a) The number of sightings of 591 species of birds in the North American Breeding Bird Survey 2003. (b) The number of addresses in the email address books of 16 881 users of a large university computer system [34]. (c) The size in acres of all wildfires occurring on US federal land between 1986 and 1996 (National Fire Occurrence Database, USDA Forest Service and Department of the Interior). Note that the horizontal axis is logarithmic in frames (a) and (c) but linear in frame (b).

Figure 1: Distributions Similar to the Power Law [1]

2. Estimation: OLS Is Biased and Inconsistent

The power law density (for a continuous distribution) or probability (for a discrete one) is

$$p(x) = Kx^{-\gamma}, \tag{1}$$

where K is a constant chosen to make the cumulative density or probability equal to one over the support of the distribution. The support is from some strictly positive value x_{min} to infinity.

If the distribution is continuous (in which case it is called the Pareto distribution) with support $[x_{min}, \infty]$, then

$$K = (\gamma - 1)x_{min}^{\gamma-1} \tag{2}$$

If the distribution is discrete (in which case it is called the Zeta or Zipf distribution) with support $\{1, 2, \ldots, \infty\}$, then

$$K = \zeta(\gamma) = \sum_{i=1}^{\infty} i^{-\gamma}. \tag{3}$$

We will work with the discrete distribution here.

Suppose we have data consisting of observations $x_1, x_2, \ldots, x_n$. This will generate an empirical probability function showing the frequency of a count, $f(z)$, with support on the possible count values $\{1, 2, \ldots, z_{max}\}$, the function that is usually graphed. One estimation approach is to use regression analysis on this generated dataset $(z, f(z))$. Most simply, one could use ordinary least squares, estimating $log(f) = \alpha - \beta log(z)$, which is equivalent to $f = \alpha z^{-\beta}$. This would be wrong. It is biased, for reasons I will explain.

Figure 2 shows the true and empirical probability functions. The horizontal axis shows $x \equiv log(count)$, where the count equals $1, ..., \infty$, so x goes from 0 to ∞. The vertical axis shows $y \equiv log(frequency(count))$. The

[1] From Newman, Figure 5.

observed frequency of a count is 0 or greater, but since $log(0) = -\infty$, any count with observed frequency of 0 is dropped from the dataset, so the remaining observed counts of y are 0 or greater. These observed counts are shown as stars in Figure 2. Ordinary least squares calculates the "OLS line" which fits those observations as well as possible.

The expected frequency of a count need not be an integer, so it can be less than 1, which means its logarithm can be negative. The large dots shows these expected freqencies. They lie on the "true line", which plots $y = \alpha - \beta x$. None of these expected frequencies are zero, so none of the counts would need to be dropped.

The observed frequency is the expected frequency plus a stochastic disturbance term, ε, so $y = \alpha - \beta x + \varepsilon$. This disturbance term has some distribution $g(\varepsilon)$. Figure 2 shows the realized values of the disturbances as vertical distances, some positive, some negative.

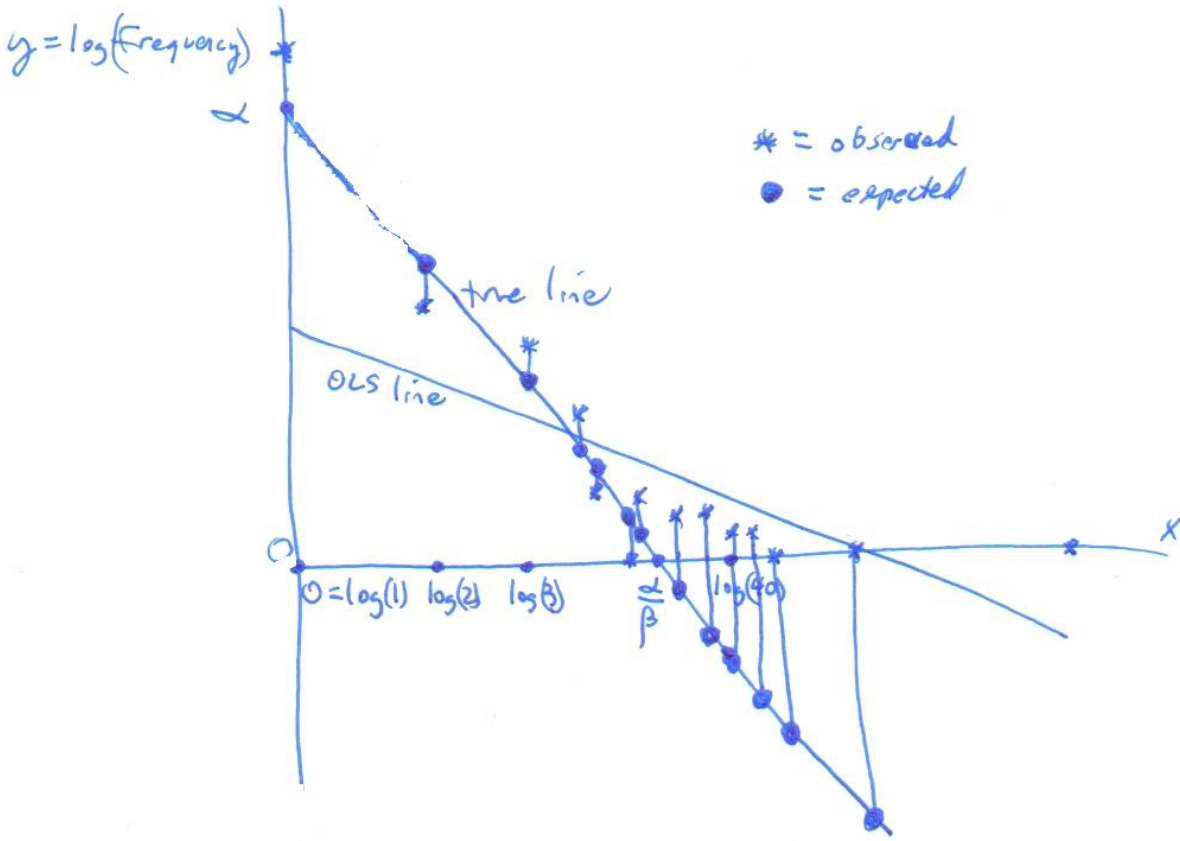

Figure 2: The Bias of Ordinary Least Squares

Let us denote the vector of T observed values of y by $\mathbf{y}$, the matrix with one column of T 1's and one column of the T observed values of X by $\mathbf{X}$, the vector of T unobserved values of ε by $\boldsymbol{\varepsilon}$, and the vector $(\alpha, -\beta)$ by $\boldsymbol{\Gamma}$. Then

$$\mathbf{y} = \mathbf{X}\boldsymbol{\Gamma} + \boldsymbol{\varepsilon}, \tag{4}$$

and the OLS estimator is

$$\hat{\boldsymbol{\Gamma}} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{y} \tag{5}$$

The expected value of the OLS estimator is

$$\begin{aligned} E\hat{\boldsymbol{\Gamma}} &= E(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{y} \\ &= E(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'(\mathbf{X}\boldsymbol{\Gamma} + \boldsymbol{\varepsilon}) \\ &= \boldsymbol{\Gamma} + E(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\boldsymbol{\varepsilon}) \end{aligned} \tag{6}$$

If $EX'\boldsymbol{\varepsilon} = 0$, OLS is unbiased. But that is not the case here, for two reasons. First, we have omitted all the observations for which $y = -\infty$. These are values for which $\boldsymbol{\varepsilon} < 0$, so the expectation of ε taken over the remaining, included, values is positive, not zero. Second, ε cannot take take negative values for large values of x, because that would result in negative values of y, and 0 is the smallest observed value of y. Thus, even the included observations have a positive expected value of ε. Both effects bias $\hat{\boldsymbol{\Gamma}}$ in a positive direction, so $\hat{\alpha}$ will be too large, and $\hat{-\beta}$ will be biased up towards zero, too small a negative number, so β will be too small. OLS is biased, and we know the sign of the bias.

As an example, consider US Supreme Court citation data. Figure 2a below shows a plot of the frequencies together with the OLS estimate, in log space and linear space.

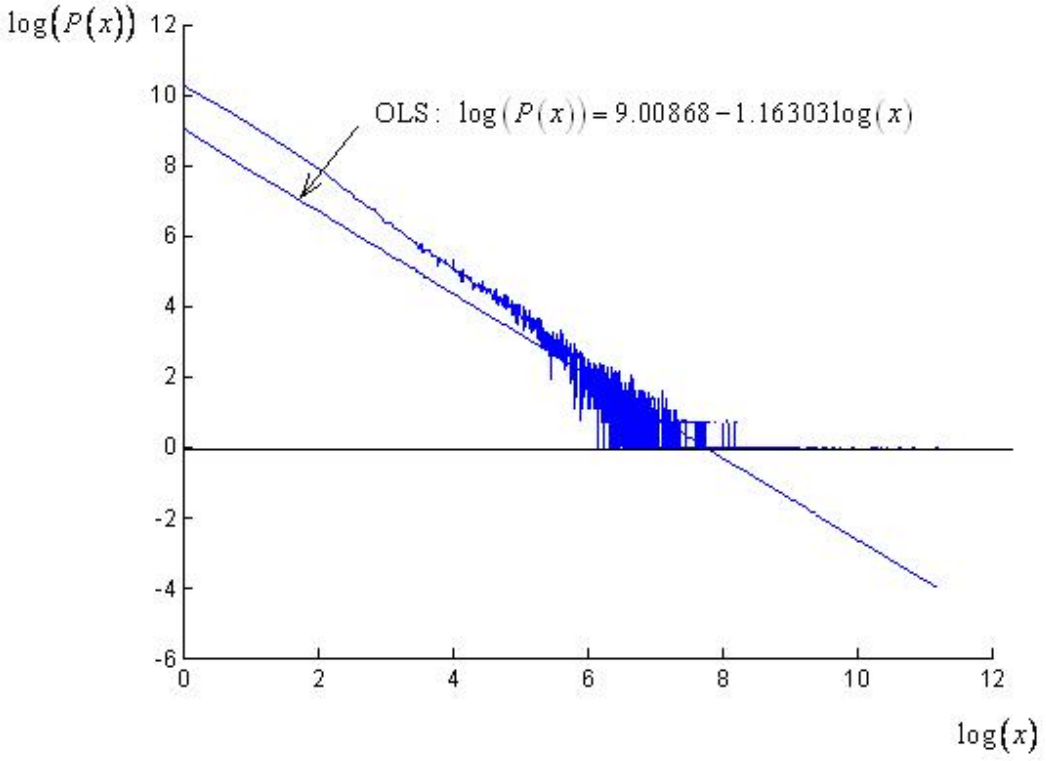

Figure 2a: U.S. Supreme Court Data with the OLS Estimate

Figure 2a: U.S. Supreme Court Data with the OLS Estimate

3. Estimation: Maximum Likelihood, Simple and Splined

The likelihood of γ being the true parameter if you observe value x_i is

$$l(\gamma|x_i) = \frac{x_i^{-\gamma}}{\zeta(\gamma)}. \tag{7}$$

Thus, the likelihood of observing the n-vector x of independent values x_i is

$$l(\gamma|x) = \Pi_{i=1}^{n} \left(\frac{x_i^{-\gamma}}{\zeta(\gamma)} \right) \tag{8}$$

It is convenient to maximize not this, but the log-likelihood (whose maximand over γ will be the same as the likelihood's), which is:

$$\begin{aligned} L(\gamma|x) &= log[l(\gamma|x)] \\ &= \sum_{i=1}^{n} [-\gamma log(x_i) - log(\zeta(\gamma))] \\ &= -\gamma \sum_{i=1}^{N} log(x_i) - N log(\zeta(\gamma)) \end{aligned} \tag{9}$$

Maximizing this with respect to γ, the first-order condition is

$$\frac{dL(\gamma|x)}{d\gamma} = -\sum_{i=1}^{n} log(x_i) - \left(\frac{\zeta'(\gamma)}{\zeta(\gamma)}\right) = 0, \tag{10}$$

so the best estimate of γ solves

$$\sum_{i=1}^{n} \frac{log(x_i)}{N} = -\left(\frac{\zeta'(\gamma)}{\zeta(\gamma)}\right) \tag{11}$$

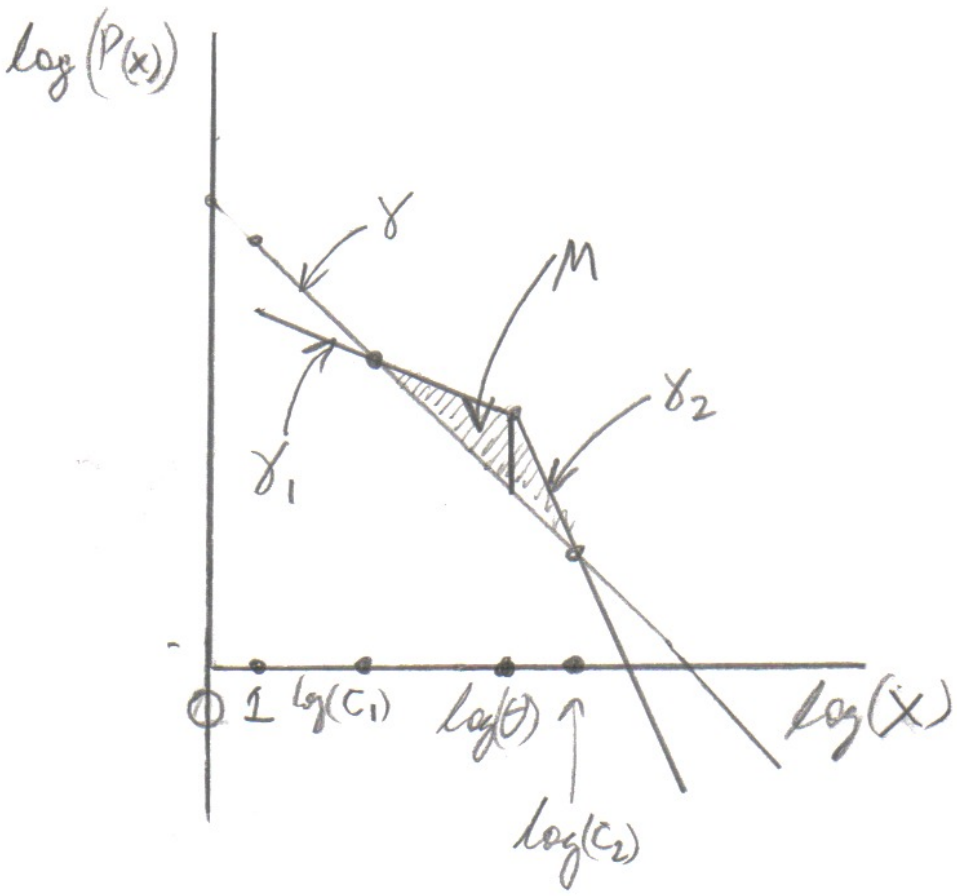

Figure 3: The Spline

Now go to having two γ's. For $x \leq \theta$, we use γ_1, and for $x \geq \theta$ we use γ_2. The two distributions must meet at θ. Thus, the probability function is

$$\begin{aligned} P(x) &= K_1 x^{-\gamma_1} \text{ if } x \leq \theta \\ &= K_2 x^{-\gamma_2} \text{ if } x > \theta \end{aligned} \tag{12}$$

Note that K_1 and K_2 do not have any necessary relation to the Zeta function. Each half of the spline does not have to be a Zeta distribution.

Without loss of generality, let us number the x_i data from smallest to biggest, so x_1 is smallest and x_N is biggest. Thus the data consists of an n-vector x such as $x = (1, 1, 1, 1, 2, 2, 3)$, which would have an empirical distribution of $f(z) =$(1:4, 2:2, 3:1). Let us define $x_{m(\theta)}$ as the last x_i that equals θ, so $x_{m(\theta)} \leq \theta < x_{m(\theta)+1}$. If $\theta = 2$, then in our example, $x_{m(\theta)} = 2$ and $m(\theta) = 6$.

There are 5 parameters in $P(x)$, but they are not all independent, just as K and γ were not earlier. Two of the five parameters can be pinned down by the following two restrictions. First, as before, the probability must sum to one.

$$\sum_{i=1}^{\theta} K_1 i^{-\gamma_1} + \sum_{i=\theta+1}^{\infty} K_2 i^{-\gamma_2} = 1 \tag{13}$$

Second, the function must be continuous, so at the cutpoint θ the two power laws intersect:

$$K_1 \theta^{-\gamma_1} = K_2 \theta^{-\gamma_2}. \tag{14}$$

These two conditions can be solved for K_1 and K_2. From (14),

$$K_2 = K_1 \theta^{\gamma_2 - \gamma_1} \tag{15}$$

From (13),

$$\sum_{i=1}^{\theta} K_1 i^{-\gamma_1} + \sum_{i=\theta+1}^{\infty} K_1 \theta^{\gamma_2 - \gamma_1} i^{-\gamma_2} = 1, \tag{16}$$

so

$$K_1 = \left(\sum_{i=1}^{\theta} i^{-\gamma_1} + \sum_{i=\theta+1}^{\infty} \theta^{\gamma_2 - \gamma_1} i^{-\gamma_2} \right)^{-1}, \tag{17}$$

Let us denote by x_j the last observation such that $x_i \leq \theta$, so $x_j \leq \theta < x_{j+1}$. The likelihood of observing the n-vector x of independent values x_i is

$$l(\gamma_1, \gamma_2, \theta | x) = \left(\Pi_{i=1}^{m(\theta)} K_1 x_i^{-\gamma_1} \right) \left(\Pi_{i=m(\theta)+1}^{N} K_2 x_i^{-\gamma_2} \right). \tag{18}$$

The log likelihood is

$$\begin{aligned} L(\gamma_1, \gamma_2, \theta|x) &= log[l(\gamma_1, \gamma_2, \theta|x)] \\ &= \sum_{i=1}^{m(\theta)} [log(K_1) - \gamma_1 log(x_i)] + \sum_{i=m(\theta)+1}^{N} [log(K_2) - \gamma_2 log(x_i)] \end{aligned} \tag{19}$$

To maximize this, use Matlab. The expressions for K_1 and K_2 contain infinite sequences, but these can be well approximated by finite sequences. As a starting point, try $\theta = N/2$ and $\gamma_1 = \gamma_2 = \gamma$ as calculated using the MLE earlier.

As an example, consider US Supreme Court citation data. Figure 3a below shows a plot of the frequencies together with the maximum likelihood estimates in log space and linear space, splined and single.

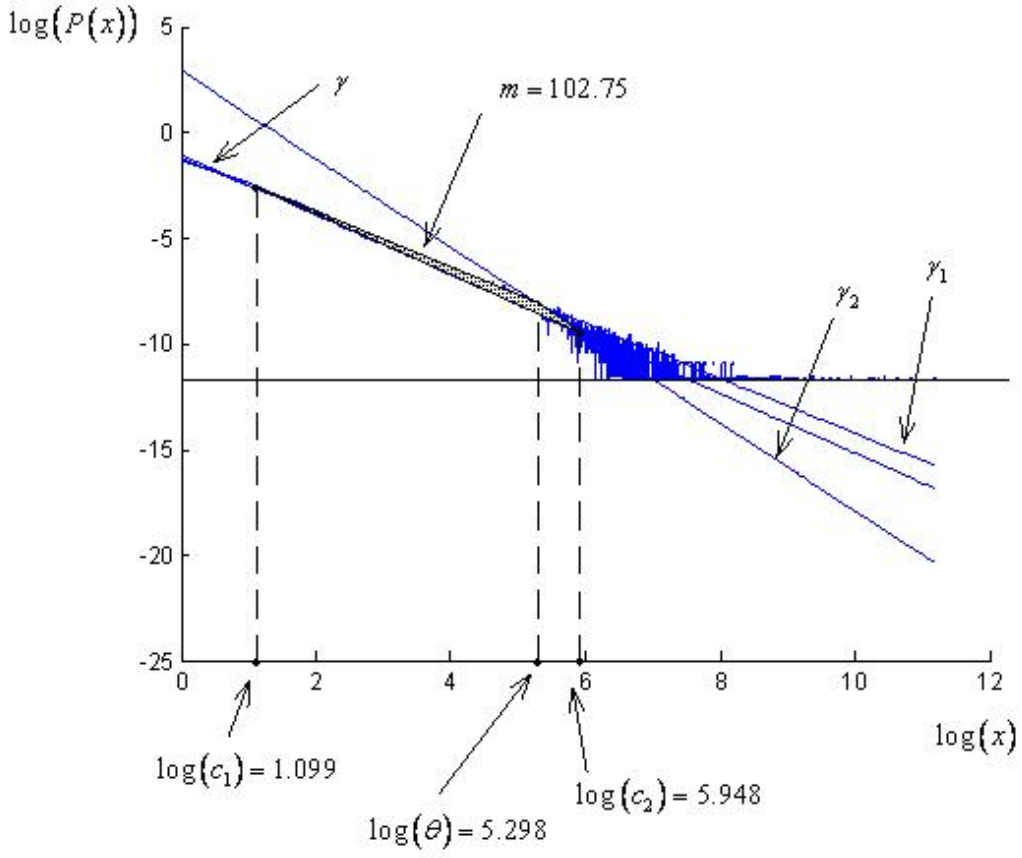

Figure 3a: U.S. Supreme Court Data with Maximum Likelihood Estimates

Figure 3a: U.S. Supreme Court Data with Maximum Likelihood Estimates

4. Testing the Estimated Curve

A first test we might do is the Kolmogorov-Smirnov test for whether it can be rejected that the single-γ power law generated the observed empirical distribution function. Figure 4 shows the critical values for this, taken from Goldstein et al.

Table 2. KS test table for power-law distributions, assuming MLE estimation.

	Quantile			
# sample	0.9	0.95	0.99	0.999
10	0.1765	0.2103	0.2835	0.3874
20	0.1257	0.1486	0.2003	0.2696
30	0.1048	0.1239	0.1627	0.2127
40	0.0920	0.1075	0.1439	0.1857
50	0.0826	0.0979	0.1281	0.1719
100	0.0580	0.0692	0.0922	0.1164
500	0.0258	0.0307	0.0412	0.0550
1000	0.0186	0.0216	0.0283	0.0358
2000	0.0129	0.0151	0.0197	0.0246
3000	0.0102	0.0118	0.0155	0.0202
4000	0.0087	0.0101	0.0131	0.0172
5000	0.0073	0.0086	0.0113	0.0147
10000	0.0059	0.0069	0.0089	0.0117
50000	0.0025	0.0034	0.0061	0.0077

Figure 4: The Kolmogorov-Smirnov Test[2]

Figure 4: The Kolmogorov-Smirnov Test[2]

If we apply the Kolmogorov-Smirnov test to the U.S. Supreme Court data, the test statistic is xxx, with a critical value of xxx at the 5 percent level.

[2]From Goldstein et al.

A second test might be a Kolmogorov-Smirnov test for whether it can be rejected that the two-γ power law generated the observed empirical distribution function. I think that might require a new table of critical values, though.

An alternative to the Kolmogorov-Smirnov test is a chi-squared test, sometimes called Pearson's Chi-Squared. To do this test, the observed data is divided into bins chosen by the researcher. The number of observations in each bin is compared to the number predicted by the single-γ or two-γ distribution. This test is sensitive to how the bins are defined. Also, an objection to it is that it would fail to detect a difference between the two distributions that occurred only within a bin— if, for example, the values $z \in [10, 15]$ were in one bin, any shape of the theoretical distribution between 10 and 15 that generates the same probability would look the same to the test.

Let us see what happens with three different bin sizes: 1 count per bin, 2 counts per bin, and 10 counts per bin. Figure 4aa, 4ab, and 4ac show the data with the expected frequency, in log space. The Chi-Squared statistics are xxx, yyy, and xxx, with critical values xxx, yyy, and xxx for the 5 percent level.

A third test is the for whether the one-γ model can be rejected in favor of the two-γ model. We will use a likelihood ratio test. See if

$$\frac{-2L(\gamma)}{L(\gamma_1, \gamma_2, \theta)} \tag{20}$$

has a value greater than Chi-Squared (2). The Neyman-Pearson Lemma says this is the most powerful test possible. Here, the test statistic is xxx, and the critical value at the 5 percent level is xxx.

A separate thing we would like to test for is "economic significance", as opposed to "statistical significance". With large amounts of data, it is easy to reject a null hypothesis, even if the difference between what is observed and the null is very small, because the measurements are so accurate. Thus, we would like some measure of how much two distributions differ, as opposed to how likely it is that they differ. I suggest using this measure, M:

$$M \equiv 1 - \left[\left(\frac{1}{2} \right) \int_{-\infty}^{\infty} \Big| f(x) - g(x) \Big| dx \right], \tag{21}$$

where$f$ and g can be either densities or probabilities.

The measure is between zero and one. If the two distributions are the same, it equals one, since $f(x) - g(x) = 0$. The smallest value is zero, since if $f(x)$ and $g(x)$ have disjoint supports,

$$1 - \left[\left(\frac{1}{2}\right)\int_{-\infty}^{\infty}\Big|f(x) - g(x)\Big|dx\right] = 1 - \left[\left(\frac{1}{2}\right)\left(\int_{-\infty}^{\infty} f(x)dx + \int_{-\infty}^{\infty} g(x)dx\right)\right] = 1 - \left[\left(\frac{1}{2}\right)(1+1)\right] \tag{22}$$

The measure can be calculated in several different ways. Note that

$$\int_{-\infty}^{\infty}\Big|f(x) - g(x)\Big|dx = \int_{x:f\geq g}[f(x) - g(x)]\,dx + \int_{x:f<g}[g(x) - f(x)]\,dx. \tag{23}$$

Since

$$\int_{-\infty}^{\infty}[f(x) - g(x)]\,dx = \int_{-\infty}^{\infty} f(x)dx - \int_{-\infty}^{\infty} g(x)dx = 1 - 1 = 0, \tag{24}$$

we know that

$$\int_{-\infty}^{\infty}[f(x) - g(x)]\,dx = \int_{x:f\geq g}[f(x) - g(x)]\,dx + \int_{x:f<g}[f(x) - g(x)]\,dx = 0, \tag{25}$$

and

$$\int_{x:f\geq g}[f(x) - g(x)]\,dx = \int_{x:f<g}[g(x) - f(x)]\,dx. \tag{26}$$

Thus, three equivalent definitions of M are

$$\begin{aligned} M &= 1 - \left[\left(\frac{1}{2}\right)\int_{-\infty}^{\infty}\Big|f(x) - g(x)\Big|dx\right] \\ &= 1 - \left[\int_{x:f\geq g}[f(x) - g(x)]\,dx\right] \\ &= 1 - \left[\int_{x:f<g}[g(x) - f(x)]\,dx.\right] \end{aligned} \tag{27}$$

We can use whichever measure is easiest to calculate.

We can use this to compare the theoretical with the empirical distribution, to get goodness of fit of the estimate. We can also use it to compare the spline to the single- γ distribution.

Do it in log space, or in regular space? Do both.

We can figure out an analytic formula for comparing the one-γ and two- γ cases. . That is easy enough for our comparison of two specified distributions. We can use the definition of M that cuts the x- space up, and look just at the difference between c_1 and c_2.

$$\begin{aligned} M &= 1 - \left(\frac{1}{2}\right)\int_{-\infty}^{\infty}\Big|f(x)-g(x)\Big|dx \\ &= 1 - \int_{c_1}^{c_2}\left[f(x)-g(x)\right]dx \\ &= 1 - \left|\int_{c_1}^{\theta}\left(f(x)-g(x)\right)dx + \int_{\theta}^{c_2}\left(f(x)-g(x)\right)dx\right| \\ &= 1 - \left|\int_{c_1}^{\theta}\left(Kx^{-\gamma}-K_1x^{-\gamma_1}\right)dx + \int_{\theta}^{c_2}\left(Kx^{-\gamma}-K_2x^{-\gamma_2}\right)dx\right| \end{aligned} \tag{28}$$

I think the two densities will cross exactly twice, at points c_1 and c_2, which I will next calculate. See Figure 2 for how this looks.

$$Kc_1^{-\gamma} = K_1c_1^{-\gamma_1}, \tag{29}$$

so

$$c_1 = \left(\frac{K}{K_1}\right)^{\gamma_1-\gamma} \tag{30}$$

and similarly,

$$c_2 = \left(\frac{K}{K_2}\right)^{\gamma_2-\gamma} \tag{31}$$

Here, I would like to do the following things using the Indiana Supreme Court data and the US Supreme Court data:

1. Graph the single-γ and two-γ and empirical distribution functions in normal space and in log space, marking c1, c2, and theta and shading the intermediate region that is used to calculate M.

2. Compute the values of M for the difference between the single-γ and two-γ and empirical distributions.

Other Measures

The Gini Coefficient for income distribution is similar. It is just for comparing a uniform density with another density, though.

http://cmc.rice.edu/docs/docs/Joh2001Mar1Symmetrizi.pdf

The KB distance is often used to compare densities. It is not a true distance– $d(f,g) \neq d(g,f)$. It is the expectation of the log likelihood ratio given that g is true, I think. It goes from zero to ifninitey.

The Akaike criterion is related somehow. It is a number for each estimate.

I can use the Akaike info criterion, for how many splines to use. it is very simple. Just $-2ln[f(y|\theta)] + 2d$, where d is the number of parameters in the model.

it msut be related to the LR test. In small smaples, this is biased, though. And it is not cnosistent.

There is not a simple explanation for why the LR ratio is chi- squaree.

Conclusions

References

M. L. Goldstein, S. A. Morris, and G. G. Yen, "Problems with fitting to the power-law distribution," *Eur. Phys. J. B*, 41: 255-258 (2004).

S. Kullback and R. A. Leibler (1951) "On Information and Sufficiency," *Annals of Mathematical Statistics,* 22(1):79-86 (March 1951).

S. Lehmann, B. Lautrup, & A. D. Jackson (2003) "Citation networks in high energy physics," *Physical Review E*, 68, 026113: 1-8 (2003)

Newman, M. E. J. "Power laws, Pareto distributions and Zipfs law"

Paul Travis Nicholls. (1989) "Bibliometric Modeling Processes and the Empirical Validity of Lotka's Law," *Journal of the American Society for Information Science,* 40(6): 379 (Nov 1989).

Wikipedia (2005) "Pareto distribution," http://en.wikipedia.org/wiki/Pareto_distribution (28 September 2005).

Wikipedia (2005) "Zeta distribution," http://en.wikipedia.org/wiki/Zeta_distribution (24 September 2005)

Andrey Feuerverger & Peter Hall (1999) "Estimating a Tail Exponent by Modelling Departure from a Pareto Distribution, " *The Annals of Statistics*, 27(2): 760-781 (April 1999).